Universitetet i Oslo
Institutt for lingvistiske og nordiske studier
Masteroppgave ved studieprogrammet IT - Språk, Logikk og Psykologi

Å trekke ut leksikalsk informasjon om tellbarheten til norske substantiver ved hjelp av korpusmetoder

Einar Stubhaug

16. juni 2005

Innhold

1 Innledning

Man kan dele substantiv opp i to grupper, **tellelige** og **utellelige** substantiv. Distinksjonen er tradisjonelt også kjent som **sortal/nonsortal, bounded/unbounded, mass/count** eller **objekt/substans**. I grammatikken gir disse dikotomiene seg utslag i at man har to ulike former for substantiv, tellelige og utellelige. De prototypisk tellelige substantivene er ord som *bil, kopp* og *gaffel*. Disse substantivene refererer til klart definerte og avgrensede fysiske objekter. Eksempler på prototypiske utellelelige substantiv er *gull* og *sand*. Det gir ingen mening å snakke om dem i flertall, de kan verken utgjøre en enhet eller deles opp i enheter, hvis man da ikke kvantifiserer eksplisitt ved hjelp av et målesubstantiv. Man kan si *4 karat gull, tre bøtter sand* osv. Prototypisk er mengdesubstantiv substans eller materie.

I denne oppgaven vil jeg først beskrive fenomenet tellbarhet mer i detalj, deretter vil jeg se på metoder for å avgjøre tellbarheten til substantiv. Jeg konsentrerer meg mest om metoder som inkluderer bruk av korpus. Jeg gjør tre eksperimenter. Det første er å lage en ikke-lærende algoritme basert på regler som jeg lager selv. Det jeg gjør der er å eksplisitt si hvordan de ulike typer substantiv vil fordele seg i korpus, ut ifra lingvistisk kunnskap. For å klassifisere et substantiv ser algoritmen på hvordan substantivet er brukt i korpus, for så å bruke reglene jeg har lagd til å gi det en klasse. I det andre eksperimentet inngår en lærende algoritme, og da er man avhengig av treningsdata. Man bruker allerede klassifiserte substantiv. For å klassifisere et nytt substantiv, ser man på hvilket av de allerede klassifiserte substantivene det ligner mest på i hvordan det forekommer i korpus.

For norsk finnes det lite ressurser på dette, og jeg har et relativt lite korpus og sparsommelig med treningsdata. Mer data og større korpus er alltid ønskelig, så tilslutt i oppgaven gjør jeg et tredje eksperiment hvor jeg forsøker å bruke internett som korpus med google som verktøy.

Alle eksempelsetningene jeg viser til er hentet fra leksikografikorpuset som Universitetet i Oslo har til rådighet.

2 Mer om tellbarhet

Typiske mengdesubstantiv er: [1]

- Substanser som væsker, gass og stoff. Eksempler på dette er *vann, vin, luft, gull, sand*. Ingen av disse substantivene kan telles. Det gir

[1] Denne listen har jeg laget selv ved å kombinere flere kilder i litteraturlisten

ingen mening å snakke om *to gull* eller *alle gullene i fjellet*. Dette har å gjøre med egenskapen til stoffet gull. Hadde gull alltid kommet i samme størrelse og form, ville man nok ha snakket om *4 guller* eller sagt noe slikt som *kan jeg betale med en gull?*. Men i vår verden gir det ingen mening å snakke om ulike enheter av gull. Det ville nok også vært tellelig hvis de menneskelige sansene hadde gitt oss inntrykk på molekylnivå, som er den eneste naturlig avgrensede enheten gull viser seg i. Hadde vi da snakket om en gullklump, ville vi kunne referere til den ved å si *alle gullene der borte*. Her begynner vi å nærme oss kjernen av hva mengdesubstantiv er, det er noe med det substantivene refererer til i verden som er udefinert, det lar seg vanskelig avgrense (det er *unbounded*) av menneskelige sanser.

- Materiale og mat. Eksempler på dette er *tre, ull, brød, kylling*. Trær finnes tellbare som fysiske objekter i verden, men materialet vi lager av trærne kodes som et mengdesubstantiv. Vi bruker samme substantiv for de to formene. Telles det, refererer vi til trærne i skogen. Telles det ikke, er det materiale det er snakk om. Det samme gjelder *kylling*, telles det er det kyllingene som sprader rundt på tunet, er det brukt som et mengdesubstantiv er det snakk om kyllingbitene i gryta.

- Tilstander, som menneskelige følelser (*angst, velvære, tristesse, kjærlighet*). Menneskelige følelser er pr definisjon utflytende (unbounded) og udefinerbare, så det er naturlig at disse blir kodet som mengdesubstantiv i språket. Det er for eksempel vanskelig å avgrense en enhet velvære.

- Egenskaper som *letthet* og *raskhet*. Det samme som over, hvordan avgrenser man *letthet*?

- Aktiviteter i form av gerunder (*løping, svømming*). Gerunder er substantiv som er avledet fra verb, og refererer til aktiviteter, ikke et fysisk objekt.

2.1 Hva gjør et konsept utellelig

Hva er det med mengdesubstantivene som gjør dem til det de er? De prototypiske mengdesubstantivene er semantisk motivert, og egenskaper ved det substantiv refererer til som gjør substantivene utellelige er bl.a.

- At enhetene er for små til at mennesker kan observere dem. Enhetene som vann og gull består av, er molekyler som vi ikke kan sanse.

- At det substantivet referer til ikke opptrer i en spesifikk størrelse og form, men varierer. Slik stoffet gull forekommer.
- At enhetene substantivet refererer til er samlet så tett at man ikke oppfatter individene, slik som sanden på en sandstrand og støvkornene i støv.
- Det er vanskelig å avgrense det substantivet referer til. Hvis det er vanskelig å definere hvor noe begynner og hvor det slutter, hvordan skal man da klare å telle det? Gode eksempler er *angst* og *velvære*.

2.2 Telling av utellelige substantiv

Mengder kan ikke telles, men de kan måles. Substansen *gull* måles i kg, er det i flytende form kan det måles i *liter*. Det kan også kvantifiseres med å sammensette *gull* med *klump* og *barre* (henholdsvis *gullklump* og *gullbarre*). *Olje* måles i *fat* og *liter*, *ved* måles i *favner*. *Øl* måles i *flasker*, *liter* og *fat*. På denne måten kan man operere med enheter også med utellelige substantiv. Mengdesubstantiv kan ikke telles fordi de ikke finnes naturlig i avgrensede enheter, men mennesker kan utmerket godt lage kunstige enheter for dem.

2.3 Konvertering

Så langt er konseptet tellelighet relativt enkelt å forholde seg til. Objektene fordeler seg i en dikotomi, hvor de enten er tellbare eller ikke tellbare. Men det viser seg at de fleste mengdesubstantiv også kan brukes tellelig, og tellelige substantiv kan også ofte brukes som mengder. Dette kaller jeg *konvertering av substantiv* (fra engelsk *coercion*).

Det er tre hovedtyper av konvertering.

- Man refererer til en spesiell type av mengdesubstantivet. *Vin*, hvis det refereres til vinen i karaffelen, er et typisk mengdesubstantiv, en væske som ikke kan telles. Men substantivet *vin* har en tellelig betydning også, men da er det snakk om ulike typer *viner*. Bruker man substantivet *vin* i flertall, refererer man til vin fra ulike steder eller vin lagd av ulike druer. Det samme gjelder olje. Stoffet som befinner seg oppi tønna kan ikke telles. Hvis olje blir telt, er det typer av olje (*olivenolje, solsikkeolje, osv*).
- Det forekommer elliptisk. Man utelater måleenheten fordi den er selvsagt. Bestiller man mengdesubstantivet *kaffe*, slik det opptrer i en beholder, ute på kafé, trenger man bare å be om "en kaffe". Hvordan kan

vi telle væsken kaffe? Et mulig svar er at vi her egentlig bestiller "en kopp kaffe", men målesubstantivet *kopp* er utelatt. I og med at både personen som serverer kaffe og den som bestiller den har en overenstemmelse om at den skal serveres i en kopp, er det ikke nødvendig å uttale det[2]. I en beskrivende episode av tv-serien *Simpsons* går Homer Simpson inn på supermarkedet og bestiller *a beer* av Apu. Apu tar fram en kagge med øl og setter på disken, hvorpå Homer sier *and a six-pack to tide me over until I can open the keg.*

- Mer sjeldent er det at prototypisk tellelige substantiv brukes som mengder. Men det er mulig, ta for eksempel uttrykk som *mye bil for penga* eller *det er mye elg i området.*

Et substantiv som kan brukes både tellbart og ikke tellbart er *fisk*. Dette belyser også problemstillingen godt. Substantivet kan brukes til å peke på et tellbart individ, en svømmende fisk, men det kan også brukes som en mengdeterm for å referere til maten vi spiser. "Se for en morsom stripete fisk!" og "min favorittmat er helt klart fisk" er eksempler på det. Men fisk kan også opptre på en tredje måte. Mange fisk har den egenskapen at de ofte opptrer i store stimer, og da har individene en tendens til å forsvinne i massen. At de i tillegg oppholder seg under vann ute av vårt synsfelt, gjør at de sjeldnere telles. Derfor kan man referere til en samling svømmende fisker som en mengde.

- Si at en dyreverner med en ekstra forkjærlighet for fisk er med en notfisker ut i båten hans. Noten fylles seg opp, og 1000 torsk blir kastet inn i lasterommet på båten. Fiskeren vil være strålende fornøyd med å ha fått så mye fisk, mens fiskeelskeren synes det er forferdelig å ha vært med på å fange så mange fisker. Ser man på det som en samling av individer eller som en masse? Dette vil variere etter personen som tolker.

- Si at du er i nærheten av et fjellvann, møter en kjentmann og du lurer på om det er bra fiske i vannet. Antageligvis vil du spørre om det er *mye fisk* i vannet. Å spørre om det er *mange fisker* der faller seg unaturlig. Er du derimot hos en kamerat og ser på akvariet hans, vil du være overrasket over hvor *mange fisker* han faktisk har der, og ikke hvor *mye fisk.*

Slik fisker opptrer i verden er det i noen situasjoner naturlig å referere til dem som en mengde, i andre sammenhenger er det naturlig å se på dem

[2]Dette er diskutert i litteraturen (Ware, 1979)

som individer. Det samme gjelder *stein*, man kan like godt si "mye stein" som "mange steiner".

Ellers kan ting konverteres av pragmatiske hensyn. Substantivet *barn* er tellbart, man kan referere til ett barn eller til fem barn om man vil. Hvis man jobber på skolefritidsordning, og det er mange barn til stede, kan man si nettopp dette, "nå var det mange barn her". Bruker man derimot det ikke tellbare perspektivet, blir betydningen annerledes. "Nå var det mye barn her"[3] gjør at man ignorerer de individuelle barna. Isteden blir barna sett på som en kollektiv skrikende og masete masse. Substantivet *barn* i mengdesforstand refererer mer til egenskapene som kjennetegner konseptet barn enn til individene selv.

2.4 Universell kverner

Ta noe tellbart, f.eks. to griser. Vi kjører det så gjennom en universell kverner (Pelletier, 1979), som kverner absolutt alt til en mos som renner ned på gulvet. Kan man da snakke om substantivet i mengdesforstand? I dette tilfellet ja, man kan si "Det er mye gris på gulvet". Dette gjelder for mange substantiv, man vil kunne si "det er mye ... på gulvet" om det, men ikke for alle. Her vil man ofte bevege seg i grenseland, og forskjellige språkbrukere vil bedømme det forskjellig. Sender man fire hammere gjennom kvernern, vil det da være mye hammer på gulvet? Det vil være mye stål og tre på gulvet, men ikke mye hammer. Dette kanskje fordi en hammer lik mange andre menneskeskapte ting, ikke lenger vil være en hammer når det knuses. Men både hester, griser og turnips vil det være mye av på gulvet. Men ta for eksempel substantivet *menneske*. Hvis jeg blir sendt gjennom kvernern, er det da mye menneske på gulvet? Grunnen til at det kan høres rart ut, er fordi mennesker for oss består av mer enn kjøtt og blod, det finnes sjel og følelser også. Det som ligger på gulvet er ikke menneske, det er rester av et menneske. En gris er for oss kun et stykke kjøtt, mens katt havner mer i gråsonen, i hvert fall for vestlige mennesker. Hadde kinesisk hatt en språklig distinksjon mellom tellelige og utellelige konsepter, ville nok *katt* kunne brukes både tellelig og utellelig.

2.5 Universell pakker

Den universelle kvernern har sitt motstykke i den universelle pakkeren (Ware, 1979). Ved pakking mener man at man tar et ubundet konsept og binder

[3] I og med at *barn* har samme form i entall og flertall, kan det diskuteres hvilken form det har her. Jeg tolker det til å være entall og at det er utellelig bruk.

det. Dette gjøres kontekstuelt, slik som i *en kaffe*, hvor mengdesubstantivet *kaffe* er bundet i en kopp. Poenget med den universelle pakkeren er at man kan alltid lage en kontekst hvor man også kan telle et "utellelig" substantiv, selv om denne konteksten kan bli temmelig søkt. Si at jeg selger jern, og bare selger det i klumper som veier nøyaktig det samme. Det ville da gitt mening å si "Jeg kjøper 4 jern".

Mengdesubstantiv kan ofte bindes med artikkelen *en*. Substantiver som *ærlighet, angst* og *misunnelse* er ikke tellbare, men man kan allikevel komme unna med setninger som under.

1. En misunnelse spiret frem i ham
2. De har en angst for noe nytt og uventet
3. Politikeren viste en forbløffende ærlighet

Dette kan synes som noe av det samme fenomenet som bestemt form entall. Mengdesubstantiv kan dukke opp i bestemt form entall, men de telles da ikke allikevel. Sier man *gullet der borte* refererer det til mengden som kan karakteriseres som gull der borte, og ikke en enhet. På samme måte i eksemplene over, substantivene telles ikke, men det blir referert til den kontekstuelt definerte mengden.[4]

2.6 Sterk og svak tellbarhet

Så de aller fleste substantiv kan konverteres. Men noen kan konverteres oftere enn andre. Som vi har sett kan substantivet *fisk* konverteres ofte, hvor det veksler mellom om de blir observert som mat, til å utgjøre en stim, eller som individuelle fisker. Det er vanskelig å klassifisere *fisk*, er det hovedsaklig tellelig eller en mengde? Andre substantiv er mer begrensede når det kommer til konvertering. Ta for eksempel *bil*, som er et klart tellbart substantiv. Det har alle kjennetegnet til tellbare substantiv, det er et avgrenset individ/objekt. Det er vanskelig å finne kontekster hvor det er mulig å bruke bil i en mengdeforstand, men det er mulig: *Mye bil for pengene* er et vanlig uttrykk som bruker bil i en mengdesforstand. I dette uttrykket er ikke referansen til substantivet avgrensede bilobjekter lenger, det blir snakk om den abstrakte ideen bil, nærmest den gamle platonske ideen.

Så langt tyder det jeg har tatt opp på at tellbarhet er semantisk motivert. Men Bond og Baldwin hevder (Baldwin & Bond, 2003b), og de er visst ikke

[4]Det kan virke som det er lettere å pakke noe med en artikkel hvis det er avsluttet og man refererer til fortiden. Man binder det ikke i **rom**, men i **tid**.

de første, at tellbarhet bør betraktes som arbitrær leksikalsk informasjon om substantivet. De gir tre gode grunner for dette:

- Et konsept kan være kodet som mengdesubstantiv i ett språk, men tellbart i et annet. Selv om det har samme referent. Til eksempel er *lightning* et mengdesubstantiv på engelsk, mens franske *éclaire* og norske *lyn* er tellbare. På samme måte er engelske *furniture* et mengdesubstantiv, mens norske *møbel* kan telles. Hvorfor teller vi noe på et språk men ikke på et annet hvis tellbarhet er et semantisk fenomen? Ord med samme referenter kan også være kodet med ulik tellbarhet i samme språk, som *things/stuff, jobs/work, clothes/clothing.* (Baldwin & Bond, 2003a)

- En undersøkelse (Imai & Gentner, 1997) viser at japanske språkbrukere oftere anser ting som mer utellelige enn engelskmenn.

- Folk fra kulturer hvis språk ikke differensierer på tellbarhet har store problemer med å lære seg tellbarheten i engelsk. Dette tyder på at konseptet om tings tellbarhet er språkavhengig, og ikke universelt semantisk.

Min konklusjon er at tellbarhet er semantisk motivert opp til et punkt. Noen ting kan ikke telles og grammatikaliseres følgelig som utellelig. Andre begreper kan være mer i gråsonen og da er det mer eller mindre arbitrært hvordan det blir kodet inn i språket.

3 Metoder for å avgjøre tellbarhet

I dette kapittelet vil jeg beskrive ulike måter å avgjøre tellbarheten til substantiver på. De forskjellige metodene har hver sine styrker og svakheter, men felles for dem alle er en del problemer som er knyttet til klassifiseringen av tellbarhet generelt.

3.1 Problemer med klassifisering

Å klassifisere er å dele inn i ulike typer klasser. Men hvilke klasser skal det deles inn i? Dette er det første problemet som dukker opp. Hvor mange klasser skal man operere i og hva er kriteriene som skal være oppfylt for at et substantiv hører hjemme i en klasse. Det man har gjort i mange tradisjonelle ordlister[5], er at man opererer med at tellelighet er standard, og at

[5]I hvert fall i det jeg har tilgang til på norsk

man markerer substantivene som ikke er tellelige. Dette er uheldig siden inndelingen ikke gjenspeiler at de fleste mengdesubstantivene kan konverteres. Og hvor går grensen for når et substantiv er utellelig eller ikke? Er det hvis det er kun marginale kontekster det kan konverteres i, eller er det hvis det aldri kan gjøres tellelig?

En løsning kan være å operere med tre klasser, en for substantiver som er tellbare (*en øks*), en for substantiv som ikke er tellbare (*juling*) og en egen klasse for substantiver som kan være både tellbare og mengder (*vin, kanin, fisk*). Men når tilhører et substantiv begge klasser? Ta for eksempel substantivet *angst*, som i likhet med andre menneskelige tilstander ikke kan bøyes i flertall. Det har alle karakteristika til mengdesubstantiv. Men allikevel er det mulig å binde det med artikkelen *en*, som i *de har en angst overfor noe nytt og uventet*. Bruken av *angst* som tellelig er helt marginal[6]. Ta et tellelig substantiv som *bil*, som i noen marginale kontekster (*mye bil for penga*) kan brukes som en mengde. Vi vil ikke at *angst* og *bil* skal havne i samme kategori.

En løsning på dette igjen er å operere med enda en klasse. Da kan vi skille mellom **tellelige** og **utellelige** substantiv, og også hybridklassene **sterkt tellelige** og **svakt tellelige** substantiv. Klassen for sterkt tellelige substantiv inneholder substantiv som i sin vanligste form er tellelige, men som også kan brukes som et mengdesubstantiv. Svakt tellelige substantiv vil da motsatt være substantiv som i sin vanligste form forekommer utellelig, men som også kan brukes tellelig. Dessverre fører dette med seg nye problemer. Ta for eksempel *vin*. Det har to ulike betydninger, typen vin og væsken vin. Er substantivet da sterkt eller svakt tellelig? Er det hvor ofte det forekommer i de ulike betydningenes som avgjør klassetilhørigheten eller er det hvilken form mennesker føler det er mest "naturlig" å klassifisere det som. Hvis det er det siste som gjelder kan det bli vanskelig for en datamaskin å avgjøre dette[7], da dette krever menneskelige bedømminger. Skal det bedømmes ut fra opptreden i korpus er korpusets representativitet avgjørende. I et aviskorpus vil det være mange vinanmeldelser, og derfor en overrepresentasjon av vin brukt tellelig. Det samme gjelder substantivet *gull*, som jeg bedømte som et rent mengdesubstantiv. Det viser seg at i aviskorpuset er gull tellelig, grunnet alle sportsreportasjene hvor *Norge tok 4 gull!*.

Løsningen jeg ser på som optimal, er å ha en ordliste som gir informasjon om substantivet er tellelig eller utellelig, og så i tillegg ha informasjon om

[6]Tidligere i oppgaven bestrider jeg at det blir telt, og hevder i stedet at det blir kontekstuelt bundet.

[7]Jeg kan se for meg en løsning hvor man forsøker å finne ut om et substantiv har en betydning "som type", "pakket", osv.

hvilke konverteringer som kan gjøres med det. *Vin* har da som oppslag at det refererer til en utellelig mengde, men kan konverteres til tellelig ved å referere til typer. *Øl* står markert som mengdesubstantiv, men det er markert at det kan telles, både som type og som pakke (*en øl*). Mange substantiv (som spesielt *fisk*) kan brukes i tellelig og utellelig form om hverandre, men da er det ikke snakk om konvertering, kun forskjellige perspektiver. Dette er ikke like interessant for ordlisten.[8]

3.2 Manuell klassifisering

Den mest åpenbare metoden er å få et menneske til å avgjøre tellbarhet manuelt. Man har gjerne en sterk tiltro til den menneskelige dømmekraft. Men dette er ikke nødvendigvis noen optimal løsning, da selv ikke fagpersoner trenger å være enige om et substantiv er tellbart eller ikke. For eksempel er de to ordlistene som Bond&Baldwin bruker, ALT/JE og COMLEX, kun enige i 93 % av substantivene (Baldwin & Bond, 2003b). Det er mange hensyn man må ta i klassifiseringen. Mange substantiv er notorisk vanskelige å klassifisere, og det vil være stor forskjell på hvilke terskler ulike språkbrukere har for å konvertere substantiv. Kombinerer man manuell klassifisering med å gi personen informasjon om hvordan substantivet han skal klassifisere opptrer i et korpus, vil dette muligens gi større enighet. Når jeg klassifiserte testdataene mine gjorde jeg flere feil, det var bl.a. flere tellelige betydninger jeg ikke tenkte på. Her ble jeg korrigert av min egen algoritme.

3.3 Klassifisering basert på ontologier

Det har vært flere forsøk på å bruke semantiske ontologier for å bestemme tellbarheten til substantiv. Et av dem er beskrevet i artikkelen "Inducing criteria for mass noun lexical mappings using the Cyc KB, and its extension to WordNet" (O'Hara, 2003).

I dette eksperimentet tas det kun hensyn til substantivers semantikk. Forfatterne bruker kunnskapsbasen Cyc, som hadde vært i utvikling i mer enn 15 år da artikkelen ble skrevet. Denne kunnskapsbasen er en ontologi som beskriver verden slik mennesker har en tendens til å konseptualisere den (O'Hara, 2003). Det er et hierarki bestående av relasjonen *isa* (er en instans av) og *genls* (er en delmengde av). For eksempel er *gaffel* er en instans av *kjøkkenredskap*, og *amfibier* er en delmengde av *dyr*. Hierarkiet består også av andre relasjoner som gir ytterligere informasjon, slik som relasjoner til andre objekter, attributter og restriksjon på bruk.

[8] Hva som er interessant å ha med i en ordliste vil det vel aldri bli enighet om.

Teknikken som brukes er beslutningstrær. Det de starter med er 250 konsepter som de har klassifisert ut i fra dikotomien mengde/tellelig. Den ene halvparten er konsepter som er i isa-kategorien, mens den andre halvparten er konsepter som er i genls-kategorien.

De henter frem alle de aktuelle ancestor-termene (konsepter som er supernoder til det aktuelle konseptet). Så induseres det over ancestor-termene og man genererer et beslutningstre med hensyn til å finne tellbarheten. En del av treet vil da se slik ut:

if OBJECTTYPE and EVENT and CREATIONEVENT **then**
if ANIMALACTIVITY **then** COUNTNOUN
else MASS NOUN
if not OBJECTTYPE and not RELATION and AGENTGENERIC
then Massnoun
if not OBJECTYPE and RELATION **then** Countnoun

Med denne framgangsmåten klarte de å gi 89 % av substantivene riktig diagnostikk. Vel og merke er det ikke substantivet som diagnostiseres, men betydningen. Konseptet *vin* har to ulike noder i ontologien, en hvor det er snakk om væsken vin, og en hvor det er snakk om typen vin. Disse vil da bli klassifisert som henholdsvis utellelig og tellelig. At et ord har to forskjellige betydninger med ulik tellbarhet er ekvivalent med at ordet blir klassifisert som et både tellbart og ikke tellbart substantiv.

Denne fremgangsmåten er interessant, men skal informasjonen kunne brukes fører den med seg et krav om at man har gode verktøy for å entydigjøre ordbetydninger. Fremgangsmåten er helt uaktuell for meg, da jeg verken har tilgang på noen ontologi eller noe slikt verktøy.

3.4 Klassifisering basert på morfologi

I noen tilfeller røper morfologien substantivets tellbarhet. Substantiv med suffikset -isme (*kapitalisme, rasisme, moralisme*) er ikke tellbare (skjønt *organisme* er et unntak). 471 substantiv i leksikografikorpuset har suffikset -isme. Substantiver som slutter på -ing er også ofte ikke tellbare (*løping, banning, kasting, juling*). Men hele 10444 substantiv slutter på -ing, og de aller fleste av dem er ikke mengdesubstantiv (*trettenåring, avdeling, nedskjæring*). Morfologi kan altså brukes, men i liten utstrekning.

3.5 Klassifisering basert på distribusjon i korpus

En metode som jeg derimot kan benytte meg av, er å se på et korpus og hvordan substantivene distribueres i det. For at denne metoden skal fungere godt, er den avhengig av at tellbare substantiv fordeler seg annerledes enn de utellbare i korpus. Jo mer annerledes, desto lettere å skille dem fra hverandre. Her kan man tenke seg mange algoritmer som kan brukes. Man kan la substantivene klynge seg sammen i grupper (*clustring*), man kan trene et klassifiseringsystem av nevrale nettverk, lage beslutningstrær osv. Men som sagt, alt dette er avhengig av at klassene har ulik distribusjon.

Her følger en oversikt over trekk som skiller tellelige og utellelige substantiv i norsk.

3.5.1 Flertall

Utellelige substantiv forekommer ikke i flertall.

- Luften kjentes ren den morgenen.
- *Han trakk luftene inn i lungene.

3.5.2 Bestemmere

Utellelige substantiv kan, i entall, bli modifisert av ubundne bestemmere som *mye, noe, masse, all* og *lite* i singular form. Det kan ikke de tellelige substantivene.

1. Det er mye *biler* i denne gaten (de fleste substantiv kan modifiseres med ubundne bestemmere i flertall).
2. *Det er mye *bil* i denne gaten.
3. Det er mye *vin* igjen i glasset.

Tellelige substantiv kan telles ved hjelp av bestemmere som *en, to, flere, alle, hver, mange, få*:

1. Han solgte tre biler på en og samme dag.
2. *Han pleide å få mange juling på den tiden.
3. *Han pleide å få mange julinger på den tiden.

3.5.3 Målesubstantiv

Man kan ikke telle mengdesubstantivene, men man kan måle dem. De lar seg kvantifiseres av målesubstantiv som kg, gram, skje, tønne, glass, kopp.

1. * Rør en sukker inn i deigen.
2. Rør en skje sukker inn i deigen.
3. Han fylte 40 liter bensin på tanken.
4. Per hentet en favn (med) ved.
5. Jeg drakk tre øl på rappen.
6. I'll have two sugars (jeg tar to sukkerbiter).

De to siste er eksempler på pakking. Her er det underforstått at man snakker om henholdsvis tre flasker øl og to sukkerbiter.

3.5.4 Ubestemt entall

Mengdesubstantiv kan utmerket godt stå i ubestemt entall alene i hodet til en substantivfrase. Tellelige substantiv gjør dette langt sjeldnere.

1. Jeg ga Ola vin.
2. *Jeg ga Ola kjøttkake.
3. Jeg ga Ola en kjøttkake.

Norsk har et mangfold av konstruksjoner som lar ubestemte substantiv stå alene uten en bestemmer (på engelsk *bare singular*).

- I mange preposisjonsfraser som *i motsetning til*, *under ledelse av*, *Egil er på vei*, *hva skal vi ha til middag?*
- Generisk lesning, hvor det ikke refereres til en unik referanse, som *det lukter menneske her*, *sykkel er det mest praktiske framkomstmiddelet i byen*, *avstand måles opp med godkjent lengdebånd*
- Verb som *få, ta, gi,* tar ofte et ubestemt objekt, som i *ta affære, gi selvstyre, få tak i*. I disse konstruksjonene mister verbene mye av sitt semantiske innhold. Uten at jeg vil regne objektet som en partikkel til verbet, går de mer sammen til en syntaktisk enhet enn f.eks. *spiser* og *kake* gjør i setningen *Per spiser kake*.

- Noen substantiv som det er udiskutablelt hvem refererer til, som nærmest er egennavn, kan stå alene. I denne kategorien finner vi ord som *Gud* og *rektor*.
- Kopulaverbene *være* og *ha*. *Han har dame*, *det er tid for omtanke*,*det er behov for vaksiner*, *vi hadde gymtime*.
- I genitivskonstruksjoner står kjernen i ubestemt entall, men her fungerer genitiven som en bestemmer. Referansen i *Einars bil* er en unik bil.
- Titler er ubestemte. *Journalist Per Edgar Kokkvold, Kaptein Sabeltann*.
- Generelt er det mye støy. Det er ikke uvanlig å la tellelige substantiv stå alene i ubestemt form entall, selv om det ville vært mer naturlig å bestemme det. Avisdelen av korpuset lager spesielt mye støy, siden artikler ganske konsekvent droppes; *Sementfabrikk eksploderte i Bagdad*

4 Datamateriale

Jeg har brukt leksikografikorpuset fra tekstlaboratoriet. Korpuset består av ca 9,2 millioner tokens. Halvparten av korpuset er skjønnlitterære verker. Resten består av en blanding bygget opp av hovedsaklig avisartikler fra de største nasjonale avisene, men også noe periodika og forskningsartikler. Dette materiale er tagget og disambiguert av Oslo-Bergen-taggeren. Av informasjonen taggeren gir meg er jeg primært interessert i hvilken ordklasse et ord tilhører (det er nødvendig å vite om en forekomst av *hest* er brukt som et adjektiv eller et substantiv), og hvis det er et substantiv, er jeg interessert i å se om det forekommer i entall eller flertall, bestemt eller ubestemt.

Som sagt disambigurerer taggeren korpus, men denne disambigueringen er ikke alltid komplett. Hvis et substantiv har samme form i ubestemt entall og flertall, vil det i mange tilfeller være umulig å disambiguere dem for en regelbasert tagger. Dette er tilfeller hvor substantivet står i ubestemt form og det ikke er noen bestemmere eller gradbøyde adjektiv foran som kan disambiguere det. Dette er kontekster hvor mengdesubstantiv forekommer hyppigere enn tellelige substantiv.

Jeg har forandret korpuset til egen bruk. Taggene Oslo-Bergen-taggeren er store og komplekse, så jeg lagde en kopi av korpuset med informasjonen som var mest relevant for meg. Så det jeg har er et korpus hvor:

- Hvert ord i korpus utgjør en linje. Ordet, ordstammen og ordklassen er delt med '£&' (f.eks. *hester £& hest £& substflertall*).
- Hvis ordet ikke er entydigjort av taggeren, gir jeg det taggen flertydig. Under klassifiseringen kan jeg da velge å ignorere disse forekomstene.
- Hvis et ord har to mulige tagger, og den ene er substantiv ubestemt entall og den andre er substantiv ubestemt flertall, gir jeg det en egen tagg, siden dette er informasjon jeg mener det er verdt å ta vare på.

Taggeren gjør noen feil, for mitt prosjekt er noen av dem alvorligere enn andre.

- *Snarere* er konsekvent tagget som et substantiv. Så når jeg ser *dette gikk mye snarere enn planlagt* og *snarere* er tagget som et substantiv i entall, tror jeg at det er et substantiv som har blitt modifisert av *mye* og følgelig vil jeg klassifisere det som et mengdesubstantiv.
- Det er et problem når det dukker opp ukjente ord i målefraser, slik som f.eks. *250 g sukker* og *200 g bacon*. Taggeren har ikke sett forkortelsen *g* før, og hopper dermed over den, virker det som. Den tagger *sukker* og *bacon* til å stå i flertall, bestemt av tallordene *200* og *250*.

For å illustrere vanskelighetene med korpus kan vi se på substantivet *sukker* og hvordan det forekommer. Det er et typisk mengdesubstantiv og bør klassifiseres som det. Det dukker opp 94 ganger i korpus. I hele 75 av disse tilfellene har det ikke blitt disambiguert og er følgelig flertydig, i alle tilfellene mellom entall og flertall. I 12 tilfeller opptrer det i bestemt form entall. I 4 tilfeller forekommer det i **flertall**, og det er i konstruksjoner hvor det har blitt tagget feil, som *250 g sukker*, hvor *g* er et ukjent ord for taggeren. Det forekommer tre ganger i ubestemt form entall uten bestemmer, her er det ikke flertydig fordi tallet har blitt røpet av et modifiserende adjektiv. En av gangene i en preposisjonsfrase, en annen etter en kvantor og den tredje etter et verb. Hvis man ser bort fra de 75 treffene som ikke har blitt disambiguert, har *sukker* en fordeling som er typisk for tellelige substantiv, siden det forekommer såpass mange ganger i flertall. Men at en så stor andel av forekomstene ikke har blitt disambiguert, skiller *sukker* fra de tellelige substantivene. Denne informasjonen bør brukes i en klassifikator.

Alt av programmering er gjort med språket **Perl** (Practical Extraction and Report Language). Dette gjelder både manipulering av korpus og konstruering av klassifiseringsalgoritmer. Som database har jeg brukt MySql versjon 4.0, og for å kommunisere med databasen har jeg brukt **DBI** (perl sin DataBase Interface modul).

5 Implementering

5.1 Testdata

I kapitlene som følger skal jeg lage og teste algoritmer. For å evaluere de forskjellige algoritmene har jeg manuelt klassifisert 128 substantiv. Meningen var at testdatasettet skulle bestå av tilfeldig valgte substantiv som har opptrådt oftere enn 20 ganger i korpus. Ved å gjøre et tilfeldig utvalg ville jeg få et datasett representativt for korpus. Men etter å ha klassifisert 80 substantiv, hvorav nesten alle var tellelige, noen mengdesubstantiv og bare to som kunne være begge deler, skiftet jeg strategi. Jeg valgte da ut 20 substantiv jeg mente tilhørte begge klasser og 20 mengdesubstantiv, de første jeg kom på. Dette betyr at testdataene mine ikke er representative for korpus, men det er viktigere å få et stort nok datasett for alle tre klassene enn det er å oppnå representativitet med korpus. Jeg er interessert i å se hvor godt algoritmene klassifiserer f.eks. substantiv som kan forekomme både tellelig og utellelig, ikke bare hvor mange den klarer å klassifisere riktig totalt.

Det jeg endte opp med er et datasett bestående av 83 tellbare, 22 som tilhører begge klasser og 23 utellbare substantiv.

Mine manuelle klassifiseringer kan sikkert bestrides, for mange av substantivene er vanskelige å bedømme.

- **Tellelige substantiv** europeer, organisasjon, forledd, retning, tilståelse, avsender, tuberkulosekontroll, innvending, evangelium, eske, slektning, styremedlem, gave, sti, betraktning, løsning, mulighet, skillelinje, bolle, kloster, næringsliv, billett, skalle, intrige, statsminister, mobiltelefon, soverom, nese, plattform, detalj, evne, venninne, helhet, salg, nål, meter, akt, paraply, erstatningskrav, bås, nett, resultatregnskap, beskyldning, lensmann, onde, interiør, skurk, tomme, kurv, fastsettelse, behandling, ordlyd, betingelse, refleks, malerinne, bunke, kostnad, faktum, hage, distribusjon, lår, tragedie, byggverk, loft, boks, skjebne, gaffel, sluse, lønning, hemmelighet, katalog, forsoning, tunnel, tue, småjente, motiv, opprør, tankegang, dyne, bukk, oppstilling.

- **Både tellelige og utellelige**

  hodepine, arbeidstid, gull, tilgivelse, løk, intuisjon, sjokolade, håndbak, ost, saft, tekst, tre, sølv, vin, olje, gris, kanin, vann, fisk, øl, kaffe, volum

- **Utellelige substantiv**

  søppel, avfall, melk, bacon, kvikksølv, beundring, spising, rettferdighet, villmark, jord, pasta, kulde, tobakk, aske, sprit, sukker, angst, solidaritet, juling, lykke, papp, lit, rang.

5.2 Trekkene jeg trekker ut

Jeg samler så inn informasjon om hvordan substantiv opptrer i korpus. Helt konkret legger jeg dette inn i en database som består av lemmaet og 15 verdier knyttet til dette. Disse verdiene er som følger:

5.2.1 Flertall

Hvor ofte substantivet forekommer i flertallsform. En flertallsforekomst regner jeg som en forekomst i tellelig forstand. Unntaket er flertallsord som *klær*, som kun forekommer i flertall, og er utellelig. Siden det gjelder veldig få ord, velger jeg å se bort fra det.

5.2.2 Telt flertall

Hvor ofte substantivet er brukt i flertall på en eksplisitt tellelig måte. Dette vil si kontekster som *mange biler*, *4 hester* osv, Merk at her vil flertallsord som *klær* ikke forekomme.

5.2.3 Bestemt entall

Hvor ofte substantivet er brukt i bestemt form entall. Strengt tatt uinteressant informasjon for meg, da substantiv i bestemt form entall ikke sier noe om telleligheten. Alle tre klasser kan forekomme her.

5.2.4 Ubestemt Subjekt

Hvor ofte substantivet forekommer i ubestemt entall først i setningen. Sannsynligvis betyr dette at subjektet er ubestemt, men det kan også være et objekt som har blitt topikalisert.

5.2.5 Preposisjonsfraser

Hvor ofte substantivet forekommer i ubestemt entall i en preposisjonsfrase.

5.2.6 Artikkel

Hvor ofte substantivet blir bestemt av en artikkel. Når det blir det kan man delvis regne substantivet som telt. Grunnen til at jeg har det i en egen kategori, er fordi mange mengdesubstantiv lar seg binde av artikler men ikke telles på andre måter.

5.2.7 Mengde

Dette er antall ganger substantivet tydelig opptrer som en mengde. Det gjelder hvis det står i entall og blir modifisert av en ubunden determinator. *de drakk mye vin*, *kan jeg få litt ost?*, *all makt til folket*. De ulike kontekstene jeg tar med:

- Substantivet blir modifisert av *mye eller litt*. Jeg har også med om kontekster hvor det blir modifisert av *mer*. I teorien er det en like god indikator, men det viser seg at terskelen for å modifisere tellelige substantiv med *mer* er relativ lav, så dette er en potensiell feilkilde.
- Modifikatoren *noe*. Problemet med *noe* er at det er et såkalt "negative polarity item", i negative kontekster kan det uten problemer modifisere tellelige substantiv.
    1. De skal bruke noe jord til ...
    2. Jeg har ikke satt noe mål ...
    3. Hopptreneren benekter at slankepress er noe problem.
- Modifikatoren *all*. Er substantivet i entall og modifisert av *all*, er det en mengde. Her dukker det opp noen unntak, som i eksempel 3 og 4 nedenfor.
    1. All makt til folket!
    2. Han mistet all troverdighet.
    3. Han hadde all grunn til å si nei.
    4. Du store all verden!
- Hvis et substantiv følger et målesubstantiv, kan man ofte bedømme tellbarhet. Jeg bruker en liste med norske målesubstantiver som var tilgjengelig på LOGON sine nettsider. Noen målesubstantiv måler i mengder, mens andre måler enheter. Jeg har selv sortert dem i de to klassene.

- En klasse med målesubstantiv som kun måler mengder: *dekar desiliter, fat, favn, flaske, glass, gram, hekto, hektogram, kilo, kilogram, gram, hektoliter, kanne, kopp, klatt, kubikkcentimeter, kubikkmeter, kvadratkilometer, kvadratmeter, liter, skvett, stykke, teskje, tønne, dråpe, fat, skive, barre.*
- En annen klasse som kun måler enheter: *antall, bande, bunke, bunt, drøss, flokk, fåtall, gruppe, haug, håndfull, klase, knippe, kompani, koppel, pakke, par, rekke, serie, skokk, snes, stabel, stim, sverm, utall, trettitall, tusental.*
- *bukett, bråte, del, favn/favner, kasse, masse, mengde, sekk, spann, tonn, bag, bøling, kurv.*
- Jeg tar til slutt i oppramsingen med målesubstantivene som ikke er nyttige til å bestemme tellbarheten:*uke, tomme, år, time, fot, fedd, dynge, centimeter, kilometer, dag, dryss, dunge, korg, kvadratalen, meter, mil, milliard, millimeter, million, minutt, mål, måned.*

5.2.8 Nøytrale bestemmere

Dette er kontekster hvor et substantiv i ubestemt entall blir bestemt av en determinator som både kan bestemme massetermer og tellelige termer, som *min* og *det.*

5.2.9 Verb

Substantivet i ubestemt entall følger et verb. Antageligvis har man da med et ubestemt objekt å gjøre.

5.2.10 Aux

Substantivet følger et predikerende verb, som *å ha* og *å være.*

5.2.11 Substantiv foran

Her har man enten med et slags målesubstantiv som ikke er på listen min å gjøre, eller en ditransitiv verbkonstruksjon med det indirekte objektet foran.

1. Han ga mannen vin.
2. * Erik solgte mannen hammer.
3. Han spiste 5 sorter sukker.

5.2.12 Tellelig entall

Substantivet står i entall og er eksplisitt tellelig. Det gjelder hvis det står bundne bestemmere som tar substantiv i entall (*hver, enhver*) foran.

5.3 Algoritme

Det jeg har nå er en database hvor alle substantivene som forekommer i leksikografikorpuset er registrert med informasjon om deres fordeling i korpus.

Tilnærmingsmåten er som følger: Jeg vil forsøke å klassifisere substantiv i tre forskjellige klasser, **tellelige** substantiv, **utellelige** substantiv og substantiv som kan være **både tellelig og utellelig**. Aller først velger jeg å kun se på de sikre treffene, dvs. treff med ubundne determinatorer foran (*mye, litt, osv*), treff i flertall, og treff i tellelig entallsbruk (*enhver, hver*). Kriteriene for at et substantiv skal ende opp i de ulike klassene er:

- For å bli klassifisert som et mengdesubstantiv må substantivet ha forekommet minst en gang etter en ubunden determinator (i mengdekontekst), og aldri i flertall eller med tellelig entallsbruk (i tellelig kontekst).
- For å bli klassifisert som et tellelig substantiv må substantivet ha forekommet minst en gang i en tellelig kontekst og ingen ganger i mengdeskontekst.
- For å bli klassifisert som et substantiv som kan være både tellelig og utellelig må det forekomme minst en gang i mengdeskontekst og en gang i tellelig kontekst.

5.3.1 Første evaluering

Med denne fremgangsmåten klassifiserer jeg de 128 manuelt klassifiserte substantivene. Dette er min baseline-klassifisator. 99 av substantivene ble klassifisert korrekt, 10 substantiv ble ikke tildelt noen klasse. 19 substantiv ble klassifisert feil.

- 88% (73) av de tellelige substantivene ble klassifisert korrekt. To typer feil ble gjort. Av de ti feilene var det 5 som ikke fikk noen tildelt noen klasse pga. manglende tellelige forekomster (*tuberkolosekontroll, næringsliv, fastsettelse, ordlyd, distribusjon*). Disse substantivene er vel og merke vanskelige også for mennesker å klassifisere. 5 tellelige substantiv ble klassifisert som substantiv som kan tilhøre begge klasser (*detalj, helhet, behandling, hemmelighet, tankegang*).

1. *detalj* er offer for "støy", da "gå i mer detalj" forekommer i korpus. Her er det nok ment "gå mer i detalj".
2. "Søken etter mer helhet" forekommer i korpus. Her er vi i gråsonen, noen vil nok synes den setningen er grei, mens andre vil reagere på den. Jeg personlig føler at her er *helhet* brukt der hvor *helhetlig* ville vært mer naturlig å bruke.
3. *Behandling* er kanskje ikke klassifisert feil, da "han fikk mye behandling da han var lagt inn" i grunn høres grei ut.
4. "I all hemmelighet" er en kollokasjon som lurer klassifikatoren min.
5. "All tankegang" forekommer en gang i korpus og er ikke noe uvanlig uttrykk. Er kanskje tankegang et mengdesubstantiv også?

- 68% (15) av substantivene som kan være både tellelig og utellelig ble klassifisert korrekt (*arbeidstid, gull, sølv, håndbak, tre, gris* og *kanin* ble klassifisert feil). *Sølv* og *gull* forekom aldri i mengdeskontekst. Men jeg observerer at taggeren veldig ofte ikke klarer å disambiguere dem, siden de er intetkjønnsord med lik entall og flertallsform. Følgelig er det mange forekomster av disse substantivene i ubestemt entall uten noen bestemmer, som tyder på at det er mengdesubstantiv. Dette bør kunne brukes i klassifiseringsalgoritmen.

- Kun 52% av mengdesubstantivene (12) av mengdesubstantivene ble klassifisert riktig. *Bacon* ble klassifisert som tellelig i stedet for utellelig. Det var ingen sikre mengdesforekomster, men i 33 tilfeller av totalt 39 har det vært flertydig mellom entall og flertall. Grunnen til den tellelige klassifiseringen er at '150 g bacon' forekommer, hvor *g* er ukjent for taggeren. *Kvikksølv, lit, rang, solidaritet og beundring* ble ikke tildelt klasser. *Lit og rang* er vanskelige å se ut fra fordelingen hvordan skal klassifiseres, det må nesten bli pga fravær av tellelighet, og at de har mange treff i gråsonen. Det samme med *solidaritet* og *beundring*. *Rettferdighet, jord, sukker, volum, angst, papp* ble også klassifisert som tellbart og havnet dermed i begge klasser. Grunnen til det er henholdsvis:

  – "... bringe til torgs "godheter" og "rettferdigheter" finnes i korpus.
  – Flertallsordet *jordene* (som i "jordene med rug på ved gårdene") står oppført med lemmaet *jord*, men bør strengt tatt ha lemmaet *jorde*. Dermed får *jord* flertallsforekomster.

- Grunnen til at *sukker* blir klassifisert slik det blir, er forklart over.
- *Angst* har, pussig nok, en flertallsforekomst, "Jeg fortalte ham om mine angster".
- *Papp* har en også en flertallsforekomst. Det viser seg at det har sneket seg inn et utsagn på svensk i den skjønnlitterære delen hvor det svenske ordet for papir, *papper*, dukker opp.

5.4 Forbedringer

Tilnærmingen ga relativt gode resultater. En del substantiv blir ikke klassifisert, og det skyldes nok størrelsen på korpus. Ellers skyldes det feil i Oslo-Bergen-taggeren og ikke minst støy i korpus. Det er to problemer som må løses:

- Mengdeskontekstene forekommer ikke hyppig nok. Jeg må ha andre kriterier i tillegg for å klassifisere substantiv som mengdesubstantiv.
- I og med at det kun kreves ett sikkert treff for medlemskap i en kategori, er klassifisatoren svært følsom for støy. Støyet i korpus har to kilder, den ene er Oslo-Bergen-taggeren selv og den andre er forfatterne av korpustekstene. Begge to gjør sine "feil" titt og ofte, så man må på en eller annen måte få gjort metoden mer robust. Det kan være en ide og høyne terskelen for hvor mange tellelige kontekster man trenger for å klassifisere noe som tellelig.

For å få gjort noe med disse to punktene må jeg se på større deler av distribusjonen. Det er ikke nok å se på de sikre treffene, men jeg må også se på de usikre. Siden utellelige substantiv oftere opptrer i ubestemt entall uten noen bestemmer foran, bør jeg bruke trekkene som det involverer, altså hvor ofte substantivet står i ubestemt entall i preposisjonsfraser, i subjektposisisjon og i objektposisjon.

Forekommer et substantiv veldig ofte i disse litt usikre kategoriene, vil jeg slutte at det er et utellelig substantiv. Men det bør kontrolleres at substantivet distribueres relativt jevnt i de tre kategoriene, for å unngå problemet med kollokasjoner. Har f.eks. et substantiv forekommet 20 ganger, hvorav 14 er i usikre kategorier, vil man ved første øyekast regne med at det er et utellelig. Ser man derimot nøyere vil man kanskje se at alle de 14 gangene følger det et verb, ta f.eks. konstruksjonen *ta affære*. *Affære* er et typisk tellbart substantiv, så man må på en eller annen måte se på distribusjonen innad i de usikre kategoriene. Mitt forslag her er å ta det mest frekvente trekket i denne gråsonen, og erstatte dets frekvens med det nest mest frekvente

trekket. På denne måten blir man kvitt problemet med de aller hyppigste kollokasjonene.

Forekommer et substantiv veldig ofte flertydig mellom entall og flertall, er det et intetkjønnsord som forekommer ubestemt. Dette er en god indikasjon på at substantivet er utellelig.

For å gjøre metoden mer robust kan det tenkes at det vil bli bedre å høyne terskelen for de sikre treffene. Dette bør være et relativt tall. Slik det er nå holder det med 1 sikkert tellelig treff for at det skal klassifiseres som tellelig, men det ville gitt liten mening å øke grensen til f.eks. 3, da **1)** substantiv med veldig høye frekvenser vil støye mer enn det uansett (forekommer et substantiv 100.000 ganger, vil det nok uansett forekomme feil 4 ganger) **2)** substantiv med veldig lave frekvenser (f.eks. de under 10) ikke vil forekomme så ofte som 4 ganger i flertall uansett tellelighet. Spørsmålet er hvor dette relative tallet skal komme fra, det bør nok ikke tas ut av løse luften. Man kan også kreve et høyere antall sikre mengdestreff. Dette er problematisk da sikre mengdestreff forekommer sjeldent. Men kanskje man kan kreve et visst antall mengdestreff samt et visst antall forekomster i ubestemt form.

5.5 Forbedret versjon

Når jeg ser et nytt substantiv, ser jeg først på de sikre treffene og spør meg: Er dette **brukt som en mengde** og er dette **brukt telt**? Algoritmen går ulike veier etter som hvordan de sikre treffene er distribuert.

- Har man sikre beviser både for tellelighet og utellelighet, klassifiserer jeg substantivet som både tellelig og utellelig
- Er det sikre beviser for tellelighet, klassifiserer jeg det som tellelig og ser om det er åpning for å klassifisere det som utellelig også. Her opererer jeg med to nye kriterier. Hvis over 30 % av forekomstene til substantivet er i ubestemt form i subjekt eller objektposisjon klassifiseres det som utellelig også. Det andre kriteriet er hvis det forekommer i mer enn 70 % av tilfellene flertydig mellom entall og flertall, da klassifiserer jeg det også som utellelig.[9]
- Er det sikre beviser for utellelighet, klassifiserer jeg det som utellelig og ser om det er åpning for å klassifisere det som tellelig også. Da ser jeg på hvor ofte det forekommer bestemt av en artikkel.

[9]Tallene 70 og 30 er tatt ut fra løse luften. Her ville det vært best å bruke en lærende algoritme for å sette tersklene

- Finnes det ingen sikre beviser for verken tellelighet eller utellelighet, er det mer sannsynlig at substantivet er utellellig enn at det er tellelig, siden beviser for tellelighet forekommer så mye hyppigere. Forekommer det mer enn 20 % av tilfellene som ubestemt objekt eller subjekt, klassifiserer jeg det som utellelig. Det samme hvis det forekommer mer enn 20 % av gangene flertydig mellom entall og flertall. Forekommer det flere enn 3 ganger bestemt av en artikkel, klassifiserer jeg det som tellelig.

Med disse forandringene blir 112 substantiv klassifisert korrekt

- 95% (79) av de tellelige substantivene ble klassifisert korrekt
- Kun 64% (14) av de som hører til begge klasser ble klassifisert korrekt
- 83% (19) av mengdesubstantivene ble klassifisert korrekt.

Jeg har sporet resultatene mine for å se hvilke av de nye kriteriene som oppfylles. Hvis substantivet ikke har noen sikre treff av verken det tellelige eller utellelige slaget, blir det klassifisert som utellelig hvis det forekommer mer enn 20% av gangene totalt ubestemt i subjekt eller objektposisjon. Dette forårsaket at *lit* og *solidaritet* ble riktig klassifisert. Hvis mer enn 20 % av forekomstene er flertydige mellom entall og flertall, skal det også klassifiseres som utellelig. Da ble *søppel, bacon* og *kvikksølv* klassifisert riktig. Substantivet klassifiseres som tellelig hvis det forekommer mer enn 3 ganger med en artikkel foran seg. Her blir *loft, forsoning* og *tankegang* riktig klassifisert. Hvis det har blitt klassifisert som tellelig, klassifiseres det som utellelig også hvis det forekommer flertydig mellom entall og flertall mer enn 70% av gangene. Dette gjør at *gull* og *sølv* ble klassifisert riktig.

Dette er gode resultater for de tellelige substantivene og mengdesubstantivene, men de som tilhører begge klassene blir fortsatt ikke helt fanget opp av algoritmen. Mye av grunnen til at det er vanskelig er at substantivene som tilhører begge klasser sjeldnere forekommer i den utellelige formen. Jeg har heller ikke funnet noen effektiv måte å skille de tellelige substantivene som av og til forekommer ubestemt uten bestemmer, fra de substantivene som kan være både tellelige og utellelige, men som ikke har noen sikre mengdestreff og generelt ikke forekommer så ofte som mengdesubstantiv.

5.5.1 Evaluering

Den symbolske algoritmen fungerer overraskende bra. I mitt, vel og merke ikke representative, testdatasett har den en presisjon på 87,5 %. Antageligvis

er presisjonen i korpus enda høyere, da tellelige substantiv nok forekommer hyppigere der enn det de gjør i testkorpuset mitt. Mange av feilene som blir gjort kan forklares, de skyldes problemer med støy/feil i korpus.

6 Bond&Baldwins eksperiment

Francis Bond og Timothy Baldwin har gjort noen eksperimenter på tellbarhet i engelsk. I artikkelen "A plethora of methods for learning English noun countability" beskrives de ulike metodene de har brukt. I denne delen av oppgaven beskriver jeg deres eksperiment, og i neste del forsøker jeg å gjøre noe av det samme med de ressursene jeg har på norsk.

6.1 Datamateriale

De bruker 2 ordlister. ALT/JE, som er utviklet for maskinoversettelse mellom japansk og engelsk, og COMLEX. ALT/JE- ordlisten har informasjon om telleligheten til 56.245 engelske substantiv, mens COMLEX har informasjon om telleligheten til 16.898 substantiv. Begge ordlistene har detaljert informasjon om tellbarheten, ALT/JE opererer med klasser som sterkt og svakt tellelig. Men enigheten ligger kun på 93% mellom de to listene. De skiller oftest lag på at ALT/JE klassifiserer substantiv som både tellelige og utellelige når COMLEX mener de er tellelige (Baldwin & Bond, 2003b).

Metoden er også avhengig av et korpus, og de har BNC (British National Corpus) til rådighet. Dette består av ca 100 millioner ord.

De fordeler substantivene i 5 kategorier.

- Tellbare
- Ikke tellbare
- Både tellbare og ikke tellbare
- Bipartitte, som f.eks (pair of) scissors. Dette er vanligere på engelsk enn på norsk, så jeg opererer ikke med denne klassen selv. Det kan late til at man på norsk går bort fra todelingen i f.eks. bukser, man kan kjøpe både *en bukse* og *et buksepar*. På engelsk kan man derimot ikke kjøpe *a pant*.
- Substantiv som bare forekommer i flertall. Dette er en klasse utellelig substantiv som huser ord som *klær*.

6.2 Algoritmen

Bond&Baldwin bruker et korpus for å lære tellbarheten til engelske substantiv. For hver forekomst av substantivet trekker man ut informasjon som beskriver det. Denne informasjonen består av følgende trekk:

- Substantivets tall når det forekommer i hodet på substantivfrasen. Er det entall eller flertall?
- Substantivets tall når det forekommer som modifikator i NP, som i *dog food*. Dette er ikke et like relevant trekk på norsk, da vi har for vane å danne et nytt substantiv bestående av modifikatoren og substantivet (*hundemat*).
- Kongruens mellom subjekt og verb. Substantivets tall når det forekommer som subjekt versus verbets tall. Dette kan heller ikke brukes da vi ikke har verbkongruens i norsk.
- Substantivets tall versus tallet til hovedsubstantivet i en konjunksjon, som i *bikkjer og søle*.
- "N1 av N2"-konstruksjoner, som i *type of dog, cup of coffee*. De registrerer nummeret til N2 i forhold til typen av N1 trekkes ut. Dette overfører jeg til målesubstantiv på norsk.
- Når substantivet er i en preposisjonsfrase, er det ubestemt eller bestemt?
- Forekomster med pronomen. Hvilke pronomen som pleier å brukes for substantivet.
- Singulære bestemmere. Hvilke singular-velgende bestemmere opptrer substantivet med? (*en bil*).
- Flertallsbestemmere. Hvilke flertallsvelgende bestemmere opptrer substantivet med? (*mange biler*)
- Ubundne bestemmere. Hvilke ubundne bestemmere (*mer, mindre*) det opptrer med og hvilket tall det da står i.

Trekkene de har valgt ligner veldig på mine[10]. De opererer derimot med et større antall trekk, og deler substantivene inn finere enn det jeg gjør. Bl.a. registrerer de **hvilken** preposisjon et substantiv følger. Dette er informasjon jeg har sett på som irrelevant for tellbarheten.

[10] Jeg valgte mine trekk uavhengig av Bond&Baldwin, før jeg leste artikkelen

6.3 Fra tekst til trekk

Korpuset må aller først preprosesseres. For å gjøre dette har Bond&Baldwin tre ulike verktøy til rådighet:

- POS tagger
- Chunker
- Dependency parser

Disse tre typene av preprosessering gir ulike typer informasjon. Grovt sett kan vi si at forsøker man å ekstrahere trekk ut fra et POS-tagget korpus, får man benyttet veldig lite informasjon. Da er man begrenset til å se på hvilket tall substantivet er i og hvilke determinatorer som eventuelt modifiserer det. Med chunking kan man se over lengre avstander og lage regler som er avhengige av kunnskap om de ulike setningsdelene. Og til sist har man dependencyparseren, som gjør at man kan se på f.eks. styring av pronomen, verbkongruens og andre mer komplekse syntaktiske trekk.

Hver type preprosessering har sin stil. Reglene som er skrevet for POS-taggeren har gjerne høy presisjon men lav funnrate, chunkeren har noe høyere funnrate, og parserens funnrate er veldig høy. Problemet til parseren er at den ofte tvinges til å ta et valg selv om den ikke er sikker, og resultatet av parsingen kan da bli feil. Resultatene fra når de bruker parseren er faktisk ikke så mye bedre enn når de bruker POS-taggeren.

6.4 Fra trekk til vektor

Når korpuset er gått gjennom, har man summert opp hvor ofte de forskjellige trekkene har forekommet for substantivene. Denne frekvensen gjør man om til 3 nye verdier:

- Korpusfrekvensen: Hvor ofte ordet forekommer med trekket **t** i forhold til antall ord (tokens) i korpus.
- Ordfrekvensen: Hvor ofte ordet forekommer med trekket **t** i forhold til hvor ofte ordet forekommer totalt (hvor vanlig trekket er for ordet).
- Trekkfrekvensen: Hvor ofte ordet forekommer med trekket **t** i forhold til alle trekk for dette ordet totalt. (Hvor prominent trekket er for substantivet).[11]

[11] Jeg tar i denne oppgaven ikke opp forskjellen mellom todimensjonale og endimensjonale trekk

6.5 Klassifisering

Nå har hvert substantiv hver sin vektor knyttet til seg. Hver vektor består av 1852 trekkverdier. Grunntanken i klassifiseringen er at substantiv som har en lik distribusjon i korpus vil ha de samme egenskapene. For å klassifisere har de et sett med treningsdata, som er allerede klassifiserte substantiv[12]. Når man da skal se på et substantiv hvis tellbarhet er ukjent, putter man det i samme klasse som treningsubstantivet med likest distribusjon. Alle klassifikatorene er basert på TiMBL versjon 4.2, et klassifikasjonssystem basert på de k-nærmeste naboer-algoritmen.

6.5.1 Valg av modell

Bond&Baldwin forsøker med to ulike typer arkitektur for klassifisatoren.

- En hvor man har en unik klassifisator. Treningsdata må da være delt inn i de ulike multiklassene, altså i tellelige, utellelige og både tellelige og utellelige substantiv. For å klassifisere et nytt substantiv finner man det substantivet det ligner mest på, og sier at de to tilhører samme klasse[13].

- Man kan operere med 4 binære klassifikatorer. Det vil si at hvert nye substantiv som skal klassifiseres går gjennom 4 ulike operasjoner. Først sjekker man om det er tellelig, så om det er utellelig, så om det er kun flertall, så om det er bipartitt. Svaret på disse spørsmålene er ja eller nei, og det er fullt mulig at et substantiv ender opp som medlem av flere klasser. Poenget er nettopp at substantiv som både er tellelige og utellelige skal bli medlem i begge de klassene. Man får 16 (4 * 4) mulige klasser (noen av disse multiklassene er høyst uvanlige). Man risikerer også at ingen av klassifikatorene vil vedkjenne seg substantivet, slik at det ender opp ikke klassifisert.

Ifølge BBs resultater er det den første modellen med de 4 binære klassifikatorene som fungerer best.

[12] Hentet fra ALT/JE og COMLEX

[13] Dette gjelder hvis man opererer med en nabo, er det flere naboer blir det litt annerledes

7 Timbl på norsk

7.1 Datasett

I denne delen forsøker jeg å bruke Bond&Baldwins metode på norsk. Her møter man mange utfordringer. Bond&Baldwin har et stort datamateriale å ta av, de har et korpus på 100 millioner ord og ordlister med fullgod informasjon om ords tellelighet. For å analysere teksten har de flere typer verktøy å ta av, både en POS-tagger, en chunker og en dependency parser.

7.1.1 Ressurser

Det jeg har å bruke, er det taggede leksikografikorpuset på ca 10 millioner ord. Jeg har verken tilgang på chunket eller på parset tekst. Derimot har jeg noe informasjon om tellbarheten. Ikke fullgod, men noe å bygge videre på er det. Problemet med dem alle er at de ikke opererer med at substantiv kan forekomme i begge former. Substantiv er enten tellelige eller ikke tellelige[14].

- Det bokmålsordboka har gjort, er å markere der den mener det er et massesubstantiv, mens den ellers ikke kommenterer noe. Bokmålsordboka har registrert 57908 av substantivene (dvs. at 57.908 substantiv har blitt tildelt et identifikasjonsnummer). 3096 er registrert med verdien 'm1e', mens 1365 har verdien 'f1e', som innebærer at de er massesubstantiv og ikke bøyes i flertall.

- En lingvist ved navn Oddmund Vestenfor har manuelt klassifisert substantivene i Oslo-Bergen-taggeren. Han har latt være å kommentere 96703 av substantivene i databasen. 19.419 av dem har han kommentert. Av disse har 13.604 fått taggen singular, som innebærer at de ikke kan bøyes i flertall. Verb som er tellelige har han bare latt være å kommentere, så når et lemma ikke har fått kommentar er det vanskelig å si om det er fordi det er tellelig eller fordi det bare ikke er kommentert.

- I IBM sin ordliste er 12.012 av de 115.980 substantivene som har et identifikasjonsnummer markert som utellelige.

Det pussige er at de tre ordlistene bare er enige med hverandre i 28 substantiv. For 3706 av substantivene klassifiserer Ibm det som utellbart,

[14]Ikke helt sant, bokmålsordboka skal ha to lemmaer hvis det er snakk om to ulike betydninger. Men så lenge ord som *sjokolade* og *vin* er oppført med bare et lemma, er ikke dette godt nok for meg.

mens bokmålsordboka og Oddmund Vestenfor ikke kommenterer. Vestenfor og Ibm er enige om at det er utellelig i 2137 av tilfellene hvor bokmålsordboka ikke kommenterer.

Et snitt mellom bokmålsordboka og Ibm gir 2155 utellelige substantiv. Snittet mellom Vestenfor og Ibm gir 2183 mengdesubstantiv. Vestenfor og Ibm har jo strengt tatt klassifisert alle substantivene, ved full enighet skulle ca. 12000 substantiv ha være klassifisert som mengdesubstantiv. Snittet mellom Vestenfor og bokmålsordboka gir kun 81 treff, uvisst av hvilken grunn. Snittet mellom de tre kildene gir som sagt 28 substantiv.

Siden jeg ikke har noen fullgod kilde for tellbarhet, må jeg forsøke å kombinere det materialet jeg har. Jeg ser på klassifiseringene som har blitt gjort av mine tre kilder og kontrollerer med mine egne korpusundersøkelser. For at et substantiv skal bli med i datasettene mine, krever jeg at det opptrer mer enn 10 ganger i korpus. Skal man lære av fordelingen må man ha noe som fordeler seg, og derfor krever jeg et visst antall treff.

- For å konstruere et datasett med utellelige substantiv, tar jeg unionen av det de tre kildene klassifiserer som utellelig og kontrollerer med mine egne korpusundersøkelser. Det vil si at hvis en av kildene klassifiserer et substantiv i korpus som et utellelig substantiv, og det **ikke forekommer i flertall** i korpus, bruker jeg substantivet i datasettet. På denne måten får jeg et sett på 1520 utellelige substantiv.

- For å konstruere et datasett med substantiv som både kan være tellelige og utellelige, har jeg liten hjelp ifra ordlistene mine. Derfor ser jeg meg nødt til å bruke mine egne korpusundersøkelser til det. Det er 209 substantiv i korpus som forekommer minst to ganger i en mengdeskontekst og minst to ganger i en tellelig kontekst. Disse substantivene lar jeg utgjøre datasettet for substantiv som kan være både tellelige og utellelige.

- For å finne de tellelige substantivene kan jeg heller ikke hvile på de tre kildene mine, da det kan være at selv om det ikke er markert som utellelig, kan opptre i mengdeskontekster. Her er jeg ute etter prototypisk tellelige substantiv. Så det jeg gjør er å ta unionen mellom mine tre kilder, for så å kontrollere i korpus. Hvis mindre enn 10% av treffene er i ubestemt form uten bestemmer, og substantivet ikke forekommer i mengdeskontekst, klassifiserer jeg det som et tellelig substantiv. På denne måten får jeg et datasett bestående av 3987 substantiver.

7.2 Multiklasse-klassifikator

Jeg bruker de samme trekkene nå som jeg samlet inn i sted, og lar Timbl forsøke å klassifisere substantivene i de 3 ulike klassene.

I eksperimentet ble 68 % av substantivene i mitt testkorpus klassifisert riktig

- 74% av de tellelige substantivene ble klassifisert riktig, altså 22 gale klassifikasjoner.
  - Tellbare substantiv som ble klassifisert som mengdesubstantiv:
    ørken, retning, avsender, tuberkulosekontroll, evangelium, løsning, statsminister, venninne, helhet, salg, tomme, fastsettelse, betingelse, malerinne, bunke, faktum, distribusjon, gaffel, tunnel.
  - Tellbare substantiv som ble klassifisert som substantiv som kan være både tellelige og utellelige:
    detalj, behandling, tankegang.
- 30% av substantivene som er både tellelige og utellelige ble klassifisert riktig, altså 14 feil.
  - *tilgivelse, løk, intuisjon, sjokolade* og *saft* ble feilaktig klassifisert som utelukkende mengdesubstantiv.
  - *arbeidstid, gull, håndbak, tre, sølv, vin, gris, kanin* og *kaffe* ble feilaktig klassifisert som tellelig.
- 80% av mengdesubstantivene ble klassifisert riktig.
  - *Bacon, sukker og lit* ble feilaktig klassifisert som tellbare.
  - *Angst og jord* ble feilaktig klassifisert som tilhørende i begge klasser.

7.2.1 Kommentarer

Dette var ikke spesielt gode resultater, de er betraktelig dårligere enn med den regelbaserte metoden. Mens man da kunne forklare de fleste av klassifikatorens feil, er det her vanskelig å si hvorfor *evangelium, løsning* og *gaffel* ble klassifisert som mengesubstantiv. Man kan tenke seg flere grunner til hvorfor det gikk galt.

- At Timbl ikke har blitt brukt slik det bør brukes. Jeg bare satte maskineriet i gang uten å sette meg særlig inn i det. Jeg bør forsøke å

justere noe på algoritmen, justere vekter og lignende. For eksempel ble det bedre resultater ved å øke antall k fra 1 til 10, da ble 91 av 128 riktig klassifisert (71%). Det er kanskje mulig å få bedre resultater ved å justere Timbl-algoritmen på andre måter i tillegg.

- At datamaterialet er for dårlig. Alt jeg har av datamateriale er jo ting som er konstruert fra ymse mer eller mindre troverdige kilder. Verst er klassifiseringen for substantivene som tilhører begge klasser, og jeg lurte litt på om det hadde med å gjøre at denne klassen ikke forekommer særlig hyppig i treningsdata. Hvis det er slik at de ulike klassenes distribusjon går noe over i hverandre, vil den lave frekvensen til substantiv som tilhører begge klasser virke inn negativt på resultatet. Forhåpentligvis er det ikke slik, det beste er om de ulike klasser fordeler seg helt ulikt.

  Får å få en pekepinn på dette, gjorde jeg et eksperiment hvor jeg innskrenket treningsdataene, her klasse var representert med kun 200 tilfeldige forekomster. Dette eksperimentet ga dårligere resultater, det var ikke flere substantiv som ble klassifisert til å være både tellelige og utellelige nå. Jeg konkluderer med at problemet er ordlistene min, og at de 200 substantivene jeg klassifiserte som tilhørende begge klasser ikke er representative nok.

7.3 To binære klassifikatorer

I stedet for å bruke en klassifikator som fordeler substantiv etter de tre klassene, bruker jeg nå to ulike klassifikatorer. En som klassifiserer substantiv etter om de kan brukes tellelige eller ikke og en som klassifiserer etter om de kan være utellelige. Substantivene vil da bli fordelt i 4 klasser, etter som det kan være tellelig men ikke utellelig, utellelig men ikke tellelig, om det kan være begge deler (begge klassifikatorene lar ordet være i sin klasse), eller om det er ingen av delene.

Jeg fikk nøyaktig samme resultat som i tilfellet multiklasse-klassifikator, men når jeg gikk fra nærmeste nabo til nærmeste 10 naboer ble 93 av de 128 substantivene klassifisert riktig, altså to flere enn med multiklasse-klassifikatoren.

8 Googling

Jeg er interessert i å se hva slags resultater jeg får ut av et enormt korpus og regler med høy presisjon. Denne muligheten har man med søkemotoren google. Jeg vet ikke nøyaktig hvor mye tekst man får tilgang på, men jeg kan anslå det med å se på frekvensen av ord. Ta for eksempel *sidestykke*, strengen forekommer på 19.200 internettsider mens lemmaet kun forekommer 12 ganger i mitt korpus. Strengen *Vin* forekommer på 1,1 millioner norske nettsider mens lemmaet vin forekommer 512 ganger i korpus.

Så størrelsen på korpuset er det ingenting å si på. Problemet er at:

- Korpuset verken er tagget eller entydiggjort. Dette kan gjøre det vanskelig å skrive regler med høy presisjon. Ta f. eks. kriteriet om at hvis et substantiv blir modifisert av *mange*, så er det tellelig. Dette er nå en potensiell feilkilde for man vet ikke om det forekommer som pronomen eller determinator.

- Mye rart skrives på internett. Det er flust av pussig grammatikk, sprø ordelinger og bruk av bindestreker som gjør at mine scripts blir lurt. Jeg tror jeg kan konkludere med at folk som illsinte skriver debattinnlegg på nyhetsgrupper ikke er så opptatt av tellbarhet.

Jeg har først registrert frekvensen av substantivet i entall modifisert av en ubunden modifikator (*mye vann, noe vann, litt vann*), så har jeg sett på frekvensen av substantivet i flertall modifisert av *mange vann, noen vann, flere vann*. Forekommer det marginalt i den ene kategorien og relativt hyppig i den andre, er det klart hvordan substantivet skal klassifiseres. Dette er tilfelle med mange av substantivene. Problemet her, på samme måte som i mine egne undersøkelser beskrevet ovenfor, er å skille mellom tellelige substantiv som også forekommer utellelig, og tellelige substantiv som forekommer utellelig som støy.

Jeg har lagd en liste over ordene i testdataene mine, sortert etter hvor ofte de forekommer foran ubundne determinatorer i forhold til hvor ofte det forekommer totalt i søkene mine.

En del av substantivene er klare. 15 av de tellelige substantivene forekommer aldri med en ubunden determinator. Ytterligere 19 tellelige substantiv forekommer 99% av gangene med de tellelige bestemmerne. Men det gjør også substantivet *kanin*, som kan tilhøre begge klasser. Tydeligvis opptrer ikke *kanin* særlig ofte utellelig. Av de substantivene som forekommer 4 ganger så ofte med tellelige enn med utellelige bestemmere, er de aller fleste tellelige substantiv. Men herfra begynner det å bli rotete, med substantiver

klassifisert som tellelige og tilhørende begge klasser i en salig blanding. Av de som forekommer 99% av gangene med ubundne bestemmere, er de aller fleste utellelige substantiv.

Problemer jeg har:

- Konstruksjoner som *mye loft- og kjellerplass*. Her er det *plass* som blir modifisert av *mye*, og ikke *loft*. Dette er vanskelig å gjøre noe med da google ignorerer bindestreker.
- Manglende disambiguering og lemmatisering er et problem. *Lit* blir klassifisert som et mengdesubstantiv siden *liter* (som er flertallsforekomsten av *lit* forekommer så ofte ("mange liter vann", "flere liter vin" osv).

Tendensen i mitt datamateriale er at tellelige substantiv kan brukes i mengdekontekster oftere enn de utellelige substantivene kan brukes tellelige. Kvernern er kraftigere enn pakkeren. Konvertering fra tellelige substantiv til utellelige forekommer i mye større grad enn jeg trodde, f.eks. er *bolle* flere ganger brukt i en mengdeskontekst (*jeg spiste noe bolle til lunch*). Et fåtall av substantivene har utelukkende treff i det ene søket, det gjelder 6 av mengdesubstantivene og 7 av de tellelige. Resten av substantivene blir brukt i begge kontekster en eller annen gang, så kan man spørre seg om det skyldes feil som at '**-**' og '**.**' ignoreres av google eller ikke.

Ved å bruke internett som korpus har skillelinjene mellom tellelige og utellelige substantiv blitt enda svakere. I testkorpuset mitt er det bare 13 substantiv som er prototypiske for sin klasse, dvs. at de bare opptrer i en type kontekster.

9 Konklusjon

I denne oppgaven har jeg gjort ulike eksperimenter. Den regelbaserte algoritmen har jeg konstruert selv, mens den lærende algoritmen er en rekonstruering av Bond&Baldwins arbeider. Dette er to ulike innfallsvinkler til klassifisering. Den store forskjellen ligger i hva kriteriene for å oppnå medlemskap i en klasse er. Med min regelbaserte algoritme er det jeg som bestemmer kriteriene for medlemskap, men med den lærende metoden blir det annerledes. Da bestemmes kriteriene av hvordan substantiv allerede antatt å tilhøre den klassen distribuerer seg i korpus.

Svakheten til den lærende algoritmen dukker opp hvis et substantiv i treningsmateriale er vurdert til å kunne forekomme både tellelig og utellelig, og det tilfeldigvis ikke forekommer f.eks. tellelig i korpus. Når man da har

et ukjent, tellelig substantiv som skal klassifiseres, risikerer man at dette substantivet er nærmeste nabo. Da blir det tellelige substantivet klassifisert galt. Den regelbaserte algoritmen får ikke dette problemet, men vil gjøre en lignende feil for substantiv som kan forekomme i begge klasser, men som ikke forekommer i den ene av dem i korpus.

Med den regelbaserte algoritmen er det korpuset som er fasiten. Blir et substantiv brukt utellelig i korpus, kan det være utellelig. Dette vil gjenspeile et språk i forandring. Hvis folk bruker *bolle* som en mengde, så kan det brukes som en mengde. Spørsmålet er hvor mange som må bruke det før man kan kalle det godt Norsk. Dette har gjennomgående vært et problem, da de sikre treffene mine gjerne kun forekommer en gang for hvert substantiv. Er dette ene treffet støy eller ikke?

Eksperimentet med internett og google kan ikke brukes til klassifisering. Til det er søkene og resultatene altfor grovkornete. Problemer med at dataene ikke er tagget og at bindestrekene ignoreres (*mye paraply-vær i Bergen*) gir mange feil. Men ved å gå gjennom treffene som google gir, ser man også at det er mange tilfeller hvor prototypisk tellelige substantiv brukes utellelig, og omvendt. Jeg har ikke noen tall på hvor stor utstrekning dette foregår i. Uansett er det tydelig at de aller fleste substantiv kan brukes både utellelig og tellelig, men som regel er en form mer naturlig å bruke.

Referanser

Johannesen, J. B., Hagen, K., & Nøklestad, A. A constraint-based tagger for norwegian. In *17th scandinavian conference of linguistics. odense working papers in language and communication* (p. 31-48).

Baldwin, T., & Bond, F. (2003a). Learning the countability of english nouns from corpus data. In *Proceedings of the 41st annual meeting of the association for computational linguistics.*

Baldwin, T., & Bond, F. (2003b). A plethora of methods for learning english countability. In *Proc of the 2003 conference on empirical methods in natural language processing.*

Bond, F., & Vatikiotis-Bateson, C. (2002). *Using an ontology to determine english countability.*

Imai, M., & Gentner, D. (1997). A crosslinguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition.*

Jackendoff, R. (1991). Parts and boundaries. *Cognition*, *41*, 9-45.

O'Hara, T. (2003). Inferring parts of speech for lexical mappings via the cyc kb, and its extension to wordnet. In *Proc. 5th international workshop on computational semantics, tilburg.*

Pelletier, F. J. (1979). Mass terms: Some philosophical problems. In (Vol. 6). Reidel.

Ware, R. X. (1979). Mass terms: Some philosophical problems. In (Vol. 6). Reidel.